Exploration of a Geometric Model of Data Fusion

Ulukbek Ibraev
School of Communication, Information and Library Studies, Rutgers University,
4 Huntington Street, New Brunswick, NJ 08901-1071. Email: ulukbek@scils.rutgers.edu

Kwong Bor Ng
Graduate School of Library and Information Studies, Queens College, City University
of New York, Rosenthal Library - Room 250, 65-30 Kissena Boulevard, Flushing,
NY 11367. Email: kbng@scils.rutgers.edu

Paul B. Kantor
School of Communication, Information and Library Studies, Rutgers University,
4 Huntington Street, New Brunswick, NJ 08901-1071. Email: kantor@scils.rutgers.edu

Some aspects of Data Fusion (DF) for Information Retrieval (IR) are explored using a set of data from the Fifth International Conference on Text Retrieval, TREC5. It has been observed from time to time that DF applied to a pair of systems or schemes for IR may yield results that are better than those of either participating scheme. It has been conjectured that this occurs only rarely, or occurs only when poor schemes are being combined, or occurs only for problems in which there are so few relevant documents that the results are probably due to statistical fluctuation. Based on a geometrical model of DF, we derive an equation for effective DF. This equation shows that in the ideal case the performance of a pair of IR schemes may be aproximated by a quadratic polynomial. We statistically test this assumption for TREC5 Routing data. Results of the regression analysis shows that our equation for the effect of DF is generally valid.

Introduction

Techniques of data fusion have been found effective in some information retrieval situations (Turtle and Croft, 1991; Fox and Shaw, 1994; Bartell, Cottrell, and Belew, 1994; Belkin, Kantor, Fox and Shaw, 1995; Lee, 1995).

Several arguments for that effectiveness can be proposed. At the most elementary level, data fusion adds an additional degree of freedom to the operation of an information retrieval scheme (that is, the degree of mixing of the two schemes included in the fusion process) and might therefore be expected to sometimes yield better results. On the other hand there is a persistent intuitive feeling that if we combine a poorer scheme with a better scheme, the result is somehow more likely to be between the two than to exceed the better scheme. However, research summarized by Ng (Ng, 1998) shows that the fused system beats the better of the two contributing systems a fair fraction of the time.

Kantor (Kantor, 1998) has proposed a kind of geometric model for understanding why this should be so. Essentially the geometric model proposes that for a given problem, in this case an information retrieval task for a specific topical quest, there is, in some large abstract space, a very best solution. Any particular real system which is at hand produces a result which can be represented by another point in that abstract space. If the system is quite good this representative point should be quite close to the ideal point. If the system is bad this representative point will be quite far away. Given a good system and a bad system we can then ask whether there is some point on the line joining them in this abstract space that is even closer to the ideal point than the good system. As the construction in Figure 1 shows, the answer to this question depends on the angle between the good system and the bad system, as seen from the ideal point in the abstract space. If this angle is small, the bad system will be farther away than the good system, and close enough to its azimuth so that the line joining them is everywhere farther from the ideal point. On the other hand, if the angles subtended by the two systems, as viewed from the optimal point, is 90 degrees then there is necessarily some point on the line connecting them which is closer to the optimal point.

The precise calculation goes as follows. We assume that in our abstract space

$$\begin{cases} P_L = P_{ideal} - \frac{e^2}{\sigma^2} \\ P_H = P_{ideal} - \frac{d^2}{\sigma^2} \\ P_B = P_{ideal} - \frac{z^2}{\sigma^2} \end{cases} \quad (1)$$

where P_L and P_H are the average precisions at top 100 documents for worse and better schemes respectively, P_B is the average precision for the best possible DF of two schemes, P_{ideal} is the average precision of the ideal scheme, and σ is some metric constant in our space. The distances between schemes in the abstract space are denoted by d, e, w, as in Figure 1. Line $P_{ideal}P_B$ has length z and is perpendicular to the line P_HP_L. The parameter α shows the weight of line P_HP_B relative to P_HP_L.

From Pythagoras' theorem, we have

$$\begin{cases} e^2 = z^2 + [w(1-\alpha)]^2 \\ d^2 = z^2 + (w\alpha)^2 \end{cases} \quad (2)$$

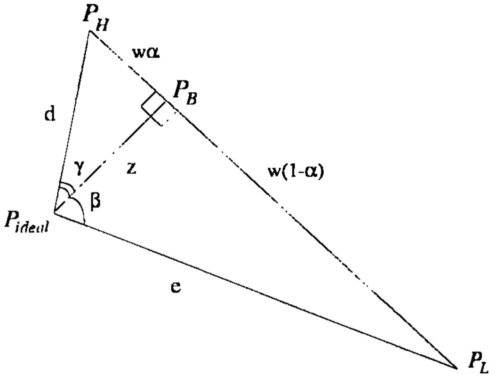

Fig. 1. The abstract space model.

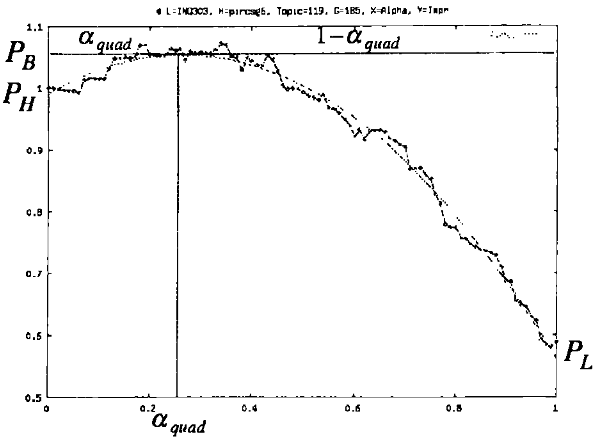

Fig. 2. Quadratic curve.

We can rewrite (2) as

$$\begin{cases} [w(1-\alpha)]^2 = e^2 - z^2 \\ (w\alpha)^2 = d^2 - z^2 \end{cases} \tag{3}$$

Dividing the first equation by the second in (3), we get

$$\frac{(1-\alpha)^2}{\alpha^2} = \frac{e^2 - z^2}{d^2 - z^2} \tag{4}$$

From (1), we have

$$\begin{cases} e^2 = P_{ideal}\sigma^2 - P_L\sigma^2 \\ d^2 = P_{ideal}\sigma^2 - P_H\sigma^2 \\ z^2 = P_{ideal}\sigma^2 - P_B\sigma^2 \end{cases} \tag{5}$$

Using (4) and (5), we get that

$$\frac{(1-\alpha)^2}{\alpha^2} = \frac{P_B - P_L}{P_B - P_H} \tag{6}$$

This simple visual presentation conveys some of the principal ideas of Kantor's geometric formulation, but is an oversimplification, in that we have not really specified the metric in this abstract space. In order for things to be as simple as suggested by our previous discussion, that metric must be chosen so that systems with equal performance (for example as measured by an average precision score in the TREC setting) must be equally distant from the optimal point. If the metric chosen can be represented by a positive quadratic form in the imagined abstract space, then an orthogonal transformation will make this true, and for purposes of presentation, we imagine that it is.

From a purely theoretical point of view, Kantor (1998) has shown that if the space is of very high dimension and systems are not correlated, it is much more likely that the angle between the two systems is close to 90 degrees than that it is close to zero. This can be understood in three dimensions by visualizing the surface of the earth. If the ideal point lies at the center of the earth and the point representing one of the two systems lies at the North Pole, then the distribution of the other system will be uniform across the surface of the earth. At angles which place this point very close to the North Pole there is relatively little surface area to be had. On the other hand at 90 degrees, the entire surface at the Equator is available for this point. Thus it is more likely, if there is no prior correlation between the two points, that the angle is close to 90 degrees, than that is close to zero. It can be shown that in spaces of substantially higher dimension this concentration in the vicinity of the Equator becomes progressively stronger. This provides some intuitive foundation for suspecting that *if* data fusion is somehow like linear combination of the representative points, it should produce results better than the best of two systems a substantial fraction of the time.

Suppose that we smooth the observed DF curve by a quadratic polynomial as in Figure 2. Here, we denoted α_{quad} as the maximum point of the quadratic polynomial given by equation $P = a\alpha^2 + b\alpha + c$ on interval [0, 1]. As we can see in Figure 2,

$$\frac{P_B - P_L}{P_B - P_H} = \frac{(1-\alpha_{quad})^2}{\alpha_{quad}^2} \tag{7}$$

However, this is the same equation as (6) with α set equal to α_{quad}. Thus, according to the geometrical model, the relationship between performance of DF and weights assigned to schemes is quadratic. The validity of this idea can be tested using statistical methods.

The present paper is an attempt to systematically explore the relation between two retrieval systems in a fusion setting, and explore characterization of the situations in which fusion does, in fact, yield results that are better than the best.

The coefficient of determination R^2

Data obtained from real experiments is usually noisy and exhibits large fluctuations. Therefore, smoothing techniques are used to interpolate the data with a well-behaved function. One of the most commonly used techniques for interpolation is the least-squares method. Given the parametric form of the function the least-squares method finds the instance of the function that best fits the real curve being interpolated.

To do this we need a measure indicating the goodness of the fit. This measure is called the *coefficient of determination* R^2. First, we need a few definitions.

Let f(x) be the function used for aproximation of the curve. Let (x_1, y_1), (x_2, y_2),...,(x_n, y_n) be n points obtained by experiment. The *residual sum of squares*, denoted by S_{res}, is defined as

$$S_{res} = \sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{8}$$

The *total sum of squares*, denoted by S_{tot}, is given by

$$S_{tot} = \sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{9}$$

where $\hat{y}$ is the average value of y.

The *coefficient of determination*, denoted by R^2, is defined as

$$R^2 = \frac{S_{tot} - S_{res}}{S_{tot}} = 1 - \frac{S_{res}}{S_{tot}} \tag{10}$$

R^2 is the proportion of variation in y that can be attributed to a relationship between x and y, in the sample. Thus, the closer this proportion is to 1, the more succesful is the relationship in explaining variation in y and therefore, the better is the fit (Devore, 1993).

The coefficient of weight α

Suppose that we have a triangle in some abstract space, as in Figure 3 . S_B and S_D are the lengths of vectors $\vec{OB}$ and $\vec{OD}$ respectively, and α is a real number such that $0 \leq \alpha \leq 1$. We want to prove that for any α, such that $0 \leq \alpha \leq 1$, the point C, built as a linear combination of S_B and S_D, lies on the line BD. It suffices to show that the vector $\vec{DC}$ is parallel to the vector $\vec{DB}$.

$$\vec{DC} = \vec{OC} - \vec{OD} = \vec{OE} + \vec{OA} - \vec{OD} = (1-\alpha)\vec{S_B} + \alpha\vec{S_D} - \vec{S_D} = (1-\alpha)(\vec{S_B} - \vec{S_D}) = (1-\alpha)\vec{DB} \tag{11}$$

The result shows that indeed $\vec{DC}$ is parallel to $\vec{DB}$, as desired. Note that, since AC is parallel to OD, the ratio

$$\frac{BC}{CD} = \frac{AB}{OA} = \frac{\alpha}{1-\alpha} \tag{12}$$

Let points P_H, P_L and P_{ideal} represent the performances of the better, worse and ideal schemes in some abstract space as before. According to the geometrical model, the best data fusion of two schemes is represented by the point P_B. Presumably, each scheme can be represented by a vector in the abstract space. Figure 4 shows the geometrical model and scheme-vectors in this space, where $\vec{S_H} = \vec{OP_H}$ and $\vec{S_L} = \vec{OP_L}$. It is clear that the best combination of two schemes is given by vector $\vec{OP_B}$. Note that the triangle OP_HP_L is similar to the triangle OBD in Figure 3. Therefore,

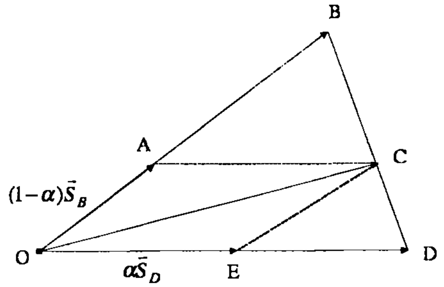

Fig. 3. A triangle in abstract space.

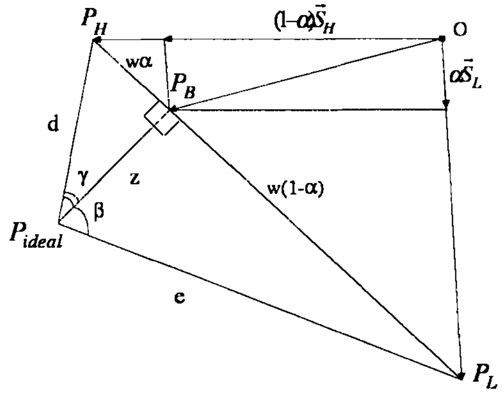

Fig. 4. Scheme vectors in the abstract space.

$$\vec{OP_B} = \alpha\vec{S_L} + (1-\alpha)\vec{S_H} \tag{13}$$

Thus, the best fusion of two schemes in the abstract space is given by a simple linear combination of the scheme vectors. The coefficient α is called the coefficient of weight of the poorer scheme or simply the weight of the poorer scheme. This simple rule might suggest that, given the relevance scores assigned by schemes, the best rule for effective data fusion is a linear combination of relevance scores with coefficients α and $(1-\alpha)$, similar to equation (13). In other words, the relevance score of a particular document is defined as

$$S_{DF} = \alpha S_L + (1-\alpha)S_H \tag{14}$$

where S_L, S_H and S_{DF} are relevance scores of the worse scheme, better scheme and data fusion respectively. Values of $\alpha > 0.5$ indicate that the worse scheme is given more weight than the better scheme, and $\alpha < 0.5$ indicate the converse.

According to the geometrical model, the closer the scheme lies to the ideal point in the abstract space, the better is its performance. Thus, the model requires that the better scheme P_H lie closer to the ideal scheme P_{ideal} than the worse scheme P_L. In other words, the distance d between P_H and P_{ideal} must be shorter than the distance e between P_L and P_{ideal}, $e > d$. Squaring both sides of the inequality and subtracting z^2, we get

$$e^2 - z^2 > d^2 - z^2 \tag{15}$$

It is easy to see in Figure 4, that this is the same as

$$(1-\alpha)^2 > \alpha^2 \tag{16}$$

or

$$\alpha < \frac{1}{2} \tag{17}$$

Thus, the geometrical model requires that the coefficient of weight assigned to the worse scheme can not exceed 0.5.

Data used

We used the output sets produced by IR schemes for the routing task of the Fifth Text Retrieval Conference (Harman, 1997) as the raw data for our study. The TREC conferences have been run as a test bed for participating groups to evaluate the performance of their IR systems based on common retrieval tasks and document collections provided by TREC. They were sponsored by National Institute of Standards and Technology (NIST) and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) as part of the TIPSTER Text Program.

The TREC5 conference centered around two main tasks, the routing task and the *ad hoc* task. In the routing task, fixed questions (topics) are asked, but new data is added to the collection. These searches are similar to those required by news clipping services and library profiling systems. In the *ad hoc* task, the new questions are asked against a static set of data. This is similar to the way that a researcher might use a library.

The participating groups could choose to do the routing task and/or the *ad hoc* task, and were asked to submit the top 1000 documents retrieved, ranked by decreasing scores of relevance, for each topic, for evaluation. The schemes were evaluated based on recall (relative recall) and precision. We used these ranked output sets as input for our data fusion experiments.

Since the output list of each of the participating IR schemes contains the relevance scores assigned by the scheme, as well as the ranked order of the 1,000 documents retrieved, we could use either the score or the rank for data fusion. In our study, we used normalized relevance scores. The normalized relevance score S is defined as:

$$S = \frac{R_{\text{orig}} - R_{\text{min}}}{R_{\text{max}} - R_{\text{min}}}$$

where R_{min} is the lowest relevance score within the top 1,000 documents retrieved for a particular topic and R_{max} is the highest relevance score within the top 1,000 documents for the same topic. Hence, the normalized relevance scores vary in the range of 0 to 1.

For the TREC 5 routing task, there were a total of 23 sets of results. We only used 22 sets of results, because one of the schemes did not comply with the conventional rule for reporting results for TREC. Each set of results had 50 topics. However, 5 topics had no relevant documents and were not used in our tests. Thus, we had 45 x 22 x 21 / 2 = 10,395 possible cases of data fusion between two IR schemes.

For each data fusion, we first normalized the relevance scores of the top 1,000 documents retrieved for a particular topic by each of the two schemes participating in the data fusion. Then we fused these two schemes according to the following formula:

$$S_{\text{new}}(\alpha) = \alpha S_{\text{worse}} + (1-\alpha) S_{\text{better}}$$

where S_{worse} and S_{better} are the normalized relevance scores assigned to a particular document by the scheme with the lower value of average precision and scheme with the higher value of average precision for a particular topic respectively, S_{new} is the new score assigned to this document, and α is the weight parameter of the data fusion which varies between 0 and 1.

To use a performance measure with an almost continuous range, we compute the average precision for the top 100 documents from the combined (fused) set. The average precision for a given topic is defined as:

$$p_{\text{ave}} = \frac{1}{G} \sum_{R} \frac{S_R}{P_R}$$

where R is a specific relevant document, S_R is the number of relevant documents up to and including the position of this relevant document, P_R is the position or rank of the relevant document in the combined set, and G is the total number of relevant documents for the given topic. We computed average precision for the top 100 documents only. Note that this score penalized a system for missing relevant documents. For example, if there are eight relevant documents for a given topic and only three of them are in the top 100 of the retrieved set at 1st, 4th, and 6th positions, then the average precision for this scheme is:

$$p_{\text{ave}} = \frac{1}{8}\left(\frac{1}{1} + \frac{2}{4} + \frac{3}{6}\right) = \frac{1}{4}$$

Results

In order to explore fusion using the data set that we have chosen, we have performed linear score fusion on all possible pairs of systems for all topics. In addition we have scanned the mixture parameter, running from $\alpha = 0$, which represents the best system alone, to $\alpha = 1$, representing the worst system alone. As α varies from 0 to 1 we can select that single mixture which produces the best possible retrieval

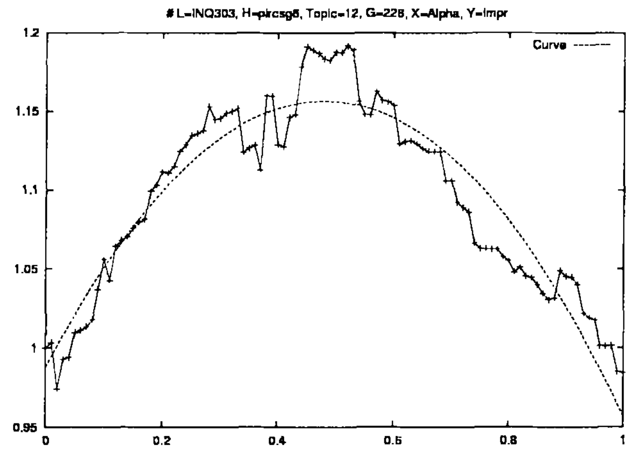

Fig. 5. Smoothing curve for INQ303, pircsg6 and topic 12.

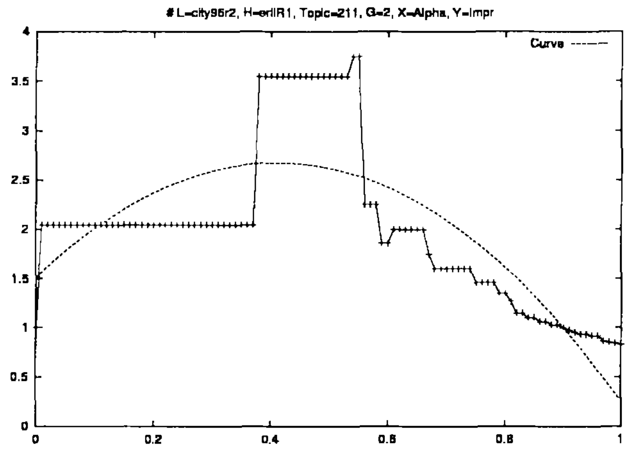

Fig. 6. Smoothing curve for city96r2, erliR1 and topic 211.

performance, as measured by the standard indicator, exact average precision.

Our computations showed that computed curves are very noisy. Statistical theory suggests that we aproximate them with some well behaved function. It was natural, given the geometrical model, to use a quadratic polynomial. Visual examination of data also suggested that the choice of a quadratic polynomial is correct. We employed the least squares method to find the best fitting polynomial for every curve. The maximum of the quadratic polynomial on the interval [0, 1] was computed and used as the best value for α, called α_{quad}.

The coefficient of determination R^2 was computed for each curve to measure the goodness of the quadratic fit. For example, Figure 5 shows the data fusion of two schemes INQ303 and pircsg6 for topic 12, which had 228 relevant documents. The X axis shows the value of α and Y axis shows the measure of improvement, P_{ave}/P_H. The computed value of R^2 for this curve was .896, indicating a good fit. Figure 6 shows the fusion curve for two schemes city96r2 and erliR1 for topic 211 which had only two relevant documents. The computed value of R^2 was .604 indicating that

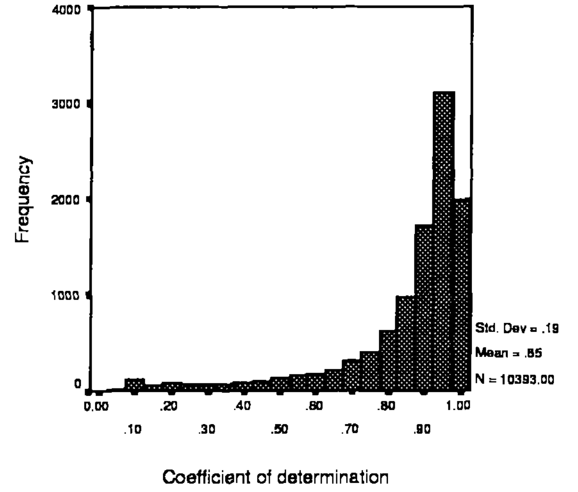

Fig. 7. Frequency distribution for R^2.

the curve can not be well smoothed by a quadratic function.

We looked for some correlation between the number of relevant documents and the goodness of fit, R^2. The results of the test showed that there is no significant correlation between the number of relevant documents and the goodness of the fit. The overall distribution of R^2 is shown in Figure 7.

Linear regression analysis

Based on visual examination of data and Ng's [1998] work, we considered the possibility that there might be a relationship between α_{quad} and P_L/P_H. Intuitively, this is attractive. If both schemes have similar performance then we would expect that efficient DF is symmetric with respect to α_{quad}, that is $\alpha_{quad} = .5$. The larger the gap between performances of better and worse schemes, the closer the value of α_{quad} to zero.

These simple observations stimulated us to conduct a linear regression analysis to check the claimed relationship. First, we conducted a linear regression between α_{quad} and P_L/P_H. The results of the analysis showed the correlation coefficient $R = .47$ and $R^2 = .22$. This shows that, indeed, these two variables correlate with each other. We also ran regression tests to see if there is a correlation between α_{quad} and a linear combination of P_L and P_H. The results for this tests were $R = .37$ and $R^2 = .14$. This means that the formula with the ratio has greater prediction power than the simpler formula saying that α_{quad} depends linearly on precisions of better and worse schemes.

Another check for validity was to examine the results of linear regression between α_{quad} and $P_L/(P_L + P_H)$. The regression analysis showed that $R = .45$ and $R^2 = .20$, almost, but not quite as good as the model based on the simple ratio. Once more, the original formula outperformed the new one.

Table 1. Results for linear regression analysis.

DV	IV	Constant	Factor	R	R^2
α_{quad}	P_L/P_H	.01	.26	.47	.22
α_{quad}	P_L, P_H	.11	.01, .00	.37	.14
α_{quad}	$\frac{P_L}{P_L+P_H}$	.00	.48	.45	.20

Table 1 shows the results of all three linear regression analyses, where DV is the dependent variable, IV is the independent variable(s), and N is the number of cases.

Discussion

This investigation confirms that data fusion in information retrieval is effective in ways that are not apparent at first sight. It is effective in some cases where improvements are achieved over the performance of very good schemes. It is effective in cases where there are large numbers of relevant documents, ruling out the explanation of "statistical fluctuations". It is effective in cases where a good scheme is improved by combining it with a relatively poor scheme.

As the outcome of this study we established a relationship between α_{quad}, the smoothed best mixing weight, and the ratio of performances of two combined schemes. We found that the predictive power of this relationship is greater than the predictive power of simpler relationship that α_{quad} depends on a linear combination of P_L and P_H. It also outperformed the model based on $P_L/(P_L + P_H)$.

However, the ratio of P_L/P_H accounts only for 22 percent of variability in α_{quad}. Thus, further research is needed to investigate other factors that might account for the rest of variability of α_{quad}.

We found that 830 cases out of the total of 10393 cases have values of $\alpha > 0.5$. These cases are counter-intuitive according to the presented geometrical model. Further research is needed to find the fundamental properties of these cases that make them different from the rest of the cases.

Acknowledgments

This research was supported in part by the US National Science Foundation Grant # IIS 9812086.

References

Bartell, B.T., Cottrell, G.W., and Belew R.K. (1994). Automatic combination of multiple ranked retrieval systems. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 173-181.

Belkin, N.J., Kantor, P.B., Fox, E., and Shaw, J. (1995). Combining the evidence of multiple query representations for information retrieval. Information Processing and Management, vol. 31, No. 3, pp.431-448.

Devore, S. and Peck, L. (1993). Statistics: The exploration and analysis of data. Belmont, CA: Duxbury Press.

Fox, E.A. and Shaw, J.A. (1994). Combination of multiple searches. Proceedings of the Third Text Retrieval Conference (TREC-3). National Institute of Standards and Technology Special Publication 500-215.

Harman, D. (1997). Overview of the Fifth Text Retrieval Conference. In D. Harman (ed.) Proceedings of the Fifth Text Retrieval Conference. Washington. DC: GPO.

Kantor, Paul. (1994) Data Fusion in Information Retrieval-Towards a Theoretical Foundation A:Vector Simulation Models. APLab Technical Report. APLab/TR-94/2

Kantor, Paul. (1998). Semantic Dimension and the Effectiveness of Linear Data Fusion Methods. In: Fifth International Conference on Artificial Intelligence and Mathematics. Jan 4-6. (unpublished; personal communication).

Lee, J.H. (1995). Combining multiple evidence from different properties of weighting schemes. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 180-188.

Lee, J.H. (1997). Analyses of multiple evidence combination. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.267-276.

Ng, K. B. and Kantor, P. B. (1998). Using scheme dissimilarity to improve data fusion for information retrieval in global libraries. In: Proceedings, 1998 Annual meeting of ASIS.

Ng, K.B. (1998). An investigation of the conditions for effective data fusion in information retrieval. Ph.D. Thesis. Graduate School of Library and Information Studies, Rutgers University.

Saracevic, T., and Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, overlap. Journal of the American Society for Information Science. Vol. 39, No. 3, pp. 197-216.

Turtle, H. and Croft, W.B. (1991) Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems. Vol. 9, No. 3, pp. 187-222.